Graph-based data integration predicts long-range regulatory interactions across the human genome

Sofie Demeyer and Tom Michoel

bioRxiv first posted online April 29, 2014
Access the most recent version at doi: http://dx.doi.org/10.1101/004622

Graph-based data integration predicts long-range regulatory interactions across the human genome

Sofie Demeyer [1] and Tom Michoel [1,*]

April 25, 2014

[1] Division of Genetics and Genomics, The Roslin Institute, The University of Edinburgh, Midlothian EH25 9RG, Scotland, United Kingdom

[*]Corresponding author: tom.michoel@roslin.ed.ac.uk

Running title: Graph-based prediction of long-range interactions

Abstract

Transcriptional regulation of gene expression is one of the main processes that affect cell diversification from a single set of genes. Regulatory proteins often interact with DNA regions located distally from the transcription start sites (TSS) of the genes. We developed a computational method that combines open chromatin and gene expression information for a large number of cell types to identify these distal regulatory elements. Our method builds correlation graphs for publicly available DNase-seq and exon array datasets with matching samples and uses graph-based methods to filter findings supported by multiple datasets and remove indirect interactions. The resulting set of interactions was validated with both anecdotal information of known long-range interactions and unbiased experimental data deduced from Hi-C and CAGE experiments. Our results provide a novel set of high-confidence candidate open chromatin regions involved in gene regulation, often located several Mb away from the TSS of their target gene.

Key words: gene regulation, long-range interactions, DNase I hypersensitive sites, gene expression, subgraph matching

1 Introduction

The central dogma of molecular biology states that genetic information flows from DNA to RNA to proteins. The rate and level at which the different genes are transcribed determine the functionality of each cell. This process is controlled by regulatory proteins binding or unbinding to/from specific DNA regions, in response to signals coming from within a cell, from neighboring cells or directly from the external environment. While some of the proteins act locally, i.e. close to the transcription start site (TSS) of the genes, others are known to bind to distal regions, even across gene boundaries (Akalin *et al.* (2009); Noonan & McCallion (2010); Ernst *et al.* (2011)). Systematically identifying these distal regulatory elements and their target genes remains one of the principal challenges of regulatory genomics.

With the help of the ENCODE consortium (ENCODE *et al.* (2011); Thurman *et al.* (2012)), large amounts of cellular data have been made publicly available. This includes both gene expression data, such as exon arrays (Kapur *et al.* (2007)) and RNA sequencing data (Nagalakshmi *et al.* (2008)), and open chromatin data, more specifically DNase-seq data. DNase sequencing (Boyle *et al.* (2008)) is a genome-wide extension of the DNase I footprinting method (Galas & Schmitz (1978)) and identifies open chromatin regions that are sensitive to cleavage by the DNase I enzyme and thus accessible to DNA-binding proteins (Wu (1980)).

Recently, linking open chromatin and gene expression information has gained attention; see Xi *et al.* (2007); Natarajan *et al.* (2012); Sheffield *et al.* (2013); Marstrand & Storey (2014). Each of these studies presents a computational method that combines DNase-seq data and gene expression data for different cell types to identify correlations between the two. However, all of these studies take only a limited number of cell types into account and are limited to open chromatin regions located in the 'proximity' of the genes (ranging between 100kb and 500kb). Nevertheless, there are known chromatin interactions between sites located several Mb away from each other (Li *et al.* (2012)).

We propose a computational method to identify interactions between open chromatin regions and genes by combining open chromatin and gene expression information for more than 100 cell types. We do not limit the area of interest around the genes, but take into account all

open chromatin regions of the whole chromosome. In this way we aim at discovering novel regulatory mechanisms, including unknown open chromatin regions interacting with genes located several Mb away.

Essentially, the proposed method obtains more accurate results by combining multiple datasets. After calculating the correlations for each dataset separately, the results are combined in a graph-based manner. This offers a distinct advantage in that the actual open chromatin and gene expression values need not to be compared directly across datasets, eliminating the need to jointly normalize the different input datasets. In addition, a similar graph-based method, used to identify open chromatin regions interacting with multiple genes, eliminates the indirect interactions.

The predicted interactions were validated with both anecdotal information of known interactions and unbiased validation sets (more specifically Hi-C (Jin *et al.* (2013)) and CAGE data (Andersson *et al.* (2014))).

2 Results

2.1 A graph-based data integration methodology

For 103 cell types, we collected DNase-seq peak data and gene expression data from the human ENCODE database (ENCODE *et al.* (2011)). DNase-seq data was collected in 100bp bins which we refer to as DNase hypersensitive sites ("DHS") (see Methods). Gene expression levels were available in the form of exon array data, which came from two different sources: 66 cell lines were collected by the University of Washington and reported exon-level data, and 37 cell lines were combined data collected by both the University of Washington and Duke University and reported gene-level data. Instead of attempting to normalise all samples to a common scale, both sets were treated as separate datasets (henceforth called the "UW" and "DukeUW" datasets) and graph-based methods were used to integrate them.

A schematic overview of the method is depicted in Figure 1. First, for each dataset, absolute Spearman correlations were calculated between all pairs of DHSs and exons/genes lying on the same chromosome. To identify the most significant interactions, an empirical null distribution was calculated from randomly permuted data and all correlations with empirical FDR values below 10% were retained (see Methods for details). This resulted in a set of DHS–gene interactions for each dataset, together with their correlation values which serve as an interaction weight or quality score. Next, a weighted edge-colored network was constructed with all DHSs, genes and exons as nodes and three types of edges: interactions from the DukeUW dataset, interactions from the UW dataset and alignment links between the different datasets (in this case mapping exons to co-located genes). This network was analyzed with the ISMA algorithm, a highly efficient subgraph matching algorithm (Demeyer *et al.* (2013)), in order to identify all subgraphs that represent an interaction occurring in both datasets (Figure 1b). For each subgraph instance, a quality score was calculated as the geometric mean of its edge weights. Next, to reconstruct unique DHS–gene interactions, we identified subgraph clusters (Michoel & Nachtergaele (2012)) (Figure 1c) and retained only those clusters (i.e. DHS–gene interactions) for which the set of exons matches with known gene transcripts (see Methods for details). The final quality score is calculated as the maximum quality score of the 3-node subgraphs in a cluster.

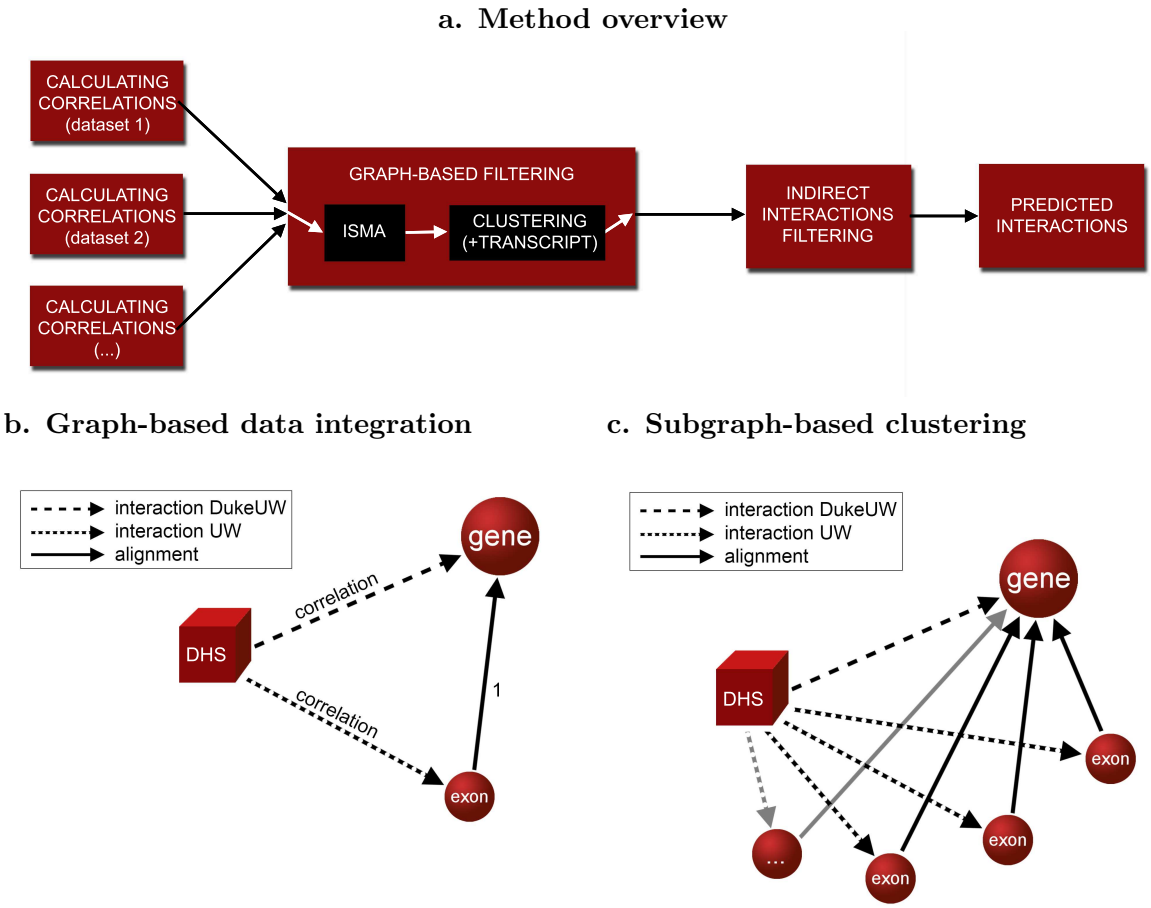


Figure 1: **Computational method overview. a.** DHS–gene regulatory interactions are predicted by a computational pipeline that combines significant correlation values calculated separately for multiple datasets (left) using graph-based data integration (middle) and indirect interaction filtering methods (right). **b.** The graph-based data integration method initially uses subgraph pattern matching to identify consistent predictions between datasets, in this case consistent significant DHS–gene correlations in the DukeUW dataset and significant DHS–exon correlations in the UW dataset. **c.** In a second step, subgraph clustering is used to group related subgraphs, in this case to detect DHS–gene correlations in the DukeUW dataset consistent with groups of DHS–exon correlations in the UW dataset, where all exons combine to form known transcripts for that gene.

Table 1 shows the number of interactions, genes/exons and DHSs found after each calculation step described above (see also Supplementary Table 2 for a complete list of the number of interactions per chromosome). As expected the graph-based filtering considerably reduces the number of predicted interactions. Hereafter, we will demonstrate that this filtering also improves the quality of the predictions.

	# 'interactions'	# genes	# DHSs
original interactions (DukeUW)	7 129 048	13 757	759 123
original interactions (UW)	890 706 846	271 780*	1 644 914
after ISMA filtering	21 234 602	6 953	247 495
		72 652*	
after clustering	2 167 676	6 953	247 495
after transcript filtering	505 716	1 833	126 543

Table 1: **Number of 'interactions' after each calculation step.** 'Interactions' represent either actual interactions between a DHS and a gene/exon (original interactions), 3-node motifs (after ISMA filtering) or single interactions between a DHS and a gene as clusters (after clustering and after transcript filtering). The genes represent either actual genes or exons (marked with a *).

2.2 Filtering indirect interactions

We calculated DHS–gene interactions from significant correlations between open chromatin areas and genes. However, some of these interactions might be of an indirect nature. Suppose, for example, a DHS is a *bona fide* regulatory element for gene A, but this gene A in turn regulates the expression level of another gene B. While correlation data alone might suggest that this DHS is also a regulatory element for gene B, it is actually a consequence of both genes interacting (Figure 2a). Our graph-based method to eliminate indirect DHS-gene correlations begins with the construction of another edge-colored network, this time containing DHS-gene interactions inferred from the previous analysis and gene-gene correlation interactions (see Methods for details). Then, we again used the ISMA algorithm (Demeyer *et al.* (2013)) to identify subgraphs that represent a single DHS interacting with two mutually correlated genes (Figure 2b(1)). Finally, we removed from these subgraphs the most weakly supported edge (Figure 2b(2–4)) and reassembled the DHS–gene interaction network from the remaining interactions, i.e. a DHS is predicted as a regulatory element for two different genes if and only if the strength of correlation between the DHS and both genes is greater than the mutual correlation between the genes.

A special case of an indirect interaction occurs when a DHS is located inside a gene body. Because expressed genes are located in open chromatin regions (Natarajan *et al.* (2012)) a large number of this type of interactions are predicted, while only some of them correspond to true regulatory interactions. We opted to remove all (looping) interactions between DHSs and co-located genes from the results. Notice that genuine enhancers that have been found within gene bodies (Arnold *et al.* (2013)) will still be predicted if the quality score of this interaction is higher than the weight of one of the two other links in the motif.

Figure 2c shows for each step of the computational pipeline the number of interactions before and after indirect interaction filtering (see Supplementary Table S3 for the numbers per chromosome). In each step of the calculations, a large number of indirect interactions are filtered out. Some of these interactions were already identified by previous filtering as the proportion of the number of interactions without the indirect ones to the total number of interactions increases with additional filtering (see Supplementary Table S2). Only 11.99% of the initial interactions remain after indirect filtering. After ISMA filtering this percentage

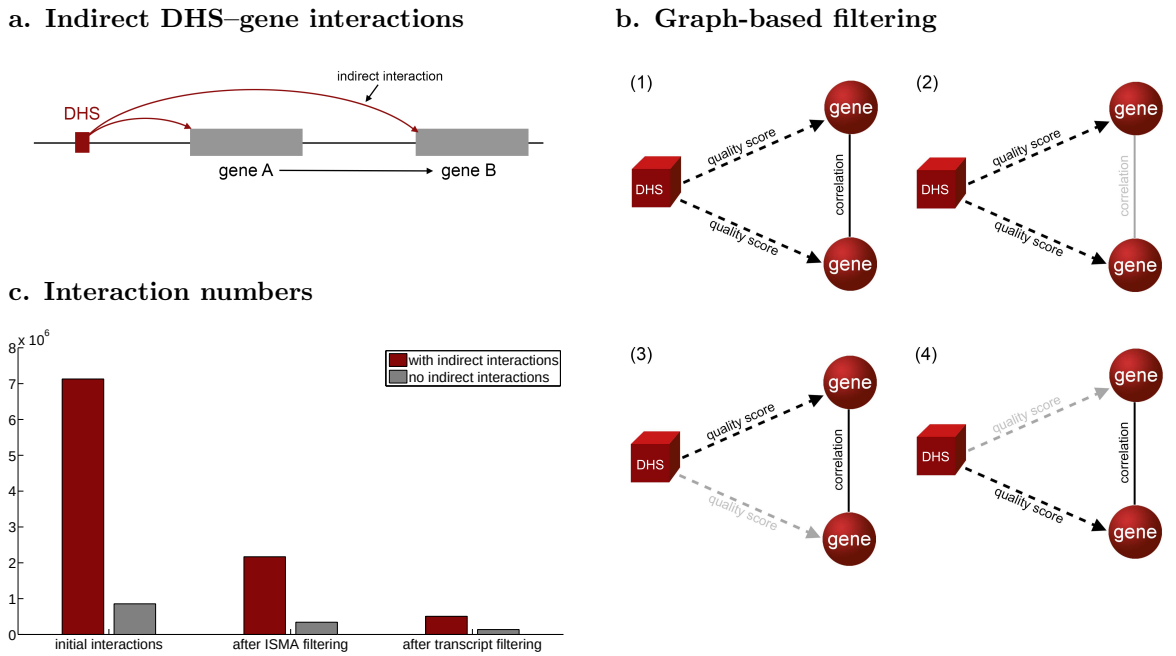


Figure 2: **Filtering indirect interactions. a.** An indirect interaction between a DHS and gene B will be inferred if gene B is regulated by gene A which in turn is regulated by the DHS. **b.** Subgraph pattern matching is used to find pairs of significant DHS–gene correlations where both genes are significantly correlated as well (1); to filter indirect interactions we remove from such subgraphs the most weakly supported edge (2–4). **c.** For each step of the pipeline the number of interactions is shown before and after indirect interaction filtering.

increases to 15.76%, and after transcript filtering to 27.17%.

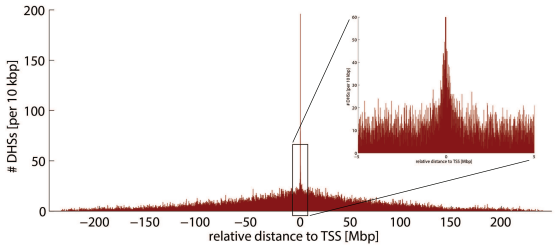
2.3 Extensive long-range interactions remain after filtering

Previous joint analyses of DNase I hypersensitivity and gene expression data have always focused on open chromatin regions located in the proximity of transcription start sites (TSSs) of genes using arbitrary distance cut-offs (varying between studies from 100–500kb) (Degner *et al.* (2012); Natarajan *et al.* (2012); Sheffield *et al.* (2013); Marstrand & Storey (2014)). Here we predicted interactions across whole chromosomes and asked whether the data supports the use of hard cut-offs proximal to the TSS.

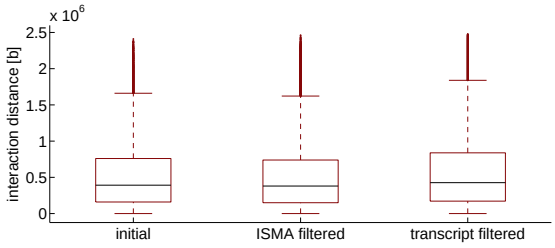
Figure 3a shows the number of interactions in function of the distance between the open chromatin areas and TSS of their interacting genes, derived from the list of interactions after all filtering steps (i.e. ISMA, transcript and indirect interactions). Although the number of interactions decreases with increasing distance, a high number of interactions occur between sites located several Mb away from each other (see also Supplementary Table S4). Similar results (see Supplementary Figure S1) were obtained for the intermediate lists of interactions, i.e. before or between the filtering steps, and no significant difference in the distributions of interaction distances was observed between the interaction lists (Figure 3b).

Next, it was investigated how the interaction quality scores are related to the distances be-

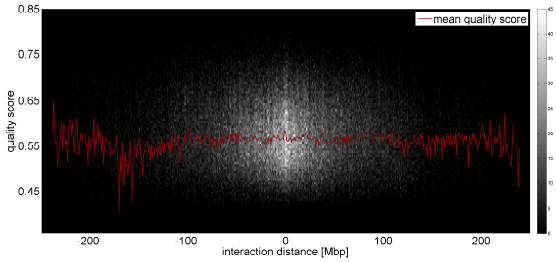
a. Number of DHSs in function of interaction distance



b. Distribution of interactions distances



c. Quality score in function of interaction distance



d. Distribution of interaction distance in function of quality score range

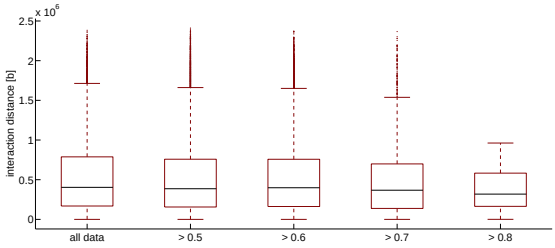


Figure 3: **Long-range interactions. a.** Number of interacting DHSs in function of the relative distance to the TSS. While there are more interacting DHSs close to the TSSs of the genes, still interactions are found between DHSs and genes located several Mb away. **b.** The distribution of the interaction distance does not change in the different filtering steps. This was also validated by the Wilcoxon rank test. **c.** The mean of quality scores (red line) remains more or less the same across the interaction distances. Again, a higher concentration of interactions was observed around the TSS. **d.** The distribution of the interaction distance does not change remarkably for different quality score cut-off values. This was confirmed with the Wilcoxon rank test.

tween the DHSs and the genes. As Figure 3c shows, the average quality score does not depend strongly on distance and interactions with high quality scores are found at all distances. Furthermore, there are no significant differences in the distance distribution at various score cut-offs, except at the very stringent threshold (>0.8) (Figure 3d). Again, similar results are obtained for the intermediate lists of interactions (see Supplementary Figure S2 and S3).

2.4 Chromosome capture data systematically validates predicted interactions

The predicted DHS–gene interactions were validated with chromosome conformation capture data. Chromosome Conformation Capture (3C) is an experimental technique to analyse the three-dimensional organization of chromosomes which reveals distal regions interacting with gene promoter regions (Dekker *et al.* (2002); Tolhuis *et al.* (2002); Wei & Zhao (2011); de Wit

& de Laat (2012)). To increase the throughput of quantifying chromosomal interactions, a number of 3C-related techniques have been developed such as Circular 3C (4C) (Zhao *et al.* (2006)), Carbon-Copy 3C (5C) (Dostie *et al.* (2006); Sanyal *et al.* (2012)) and Hi-C (Lieberman-Aiden *et al.* (2009)). When a DHS is predicted to (functionally) interact with a certain gene by our method and it is found by chromosome capture data to lie in a chromosomal region that physically interacts with that gene's promoter, we considered this a validation of the predicted interaction.

2.4.1 Genome-wide 3C (Hi-C) validation

Genome-wide Hi-C data allows for the most systematic validation of predicted DHS–gene interactions. Here we used a Hi-C dataset (the 'gold standard') from IMR90 (primary human fibroblast) cells which contained 57,059 interactions between 49,394 regions (1-20 kbp) centered on the *cis*-elements annotated in the IMR90 cell genome and the promoter regions of 9181 genes with a FDR of 10%, reporting only interactions within a 2 Mb distance (Jin *et al.* (2013)). Following established protocols from the network inference field (Stolovitzky *et al.* (2009)), we only considered DHS–gene predictions within the gold standard space (i.e. the set of all possible interactions between DHSs and genes from the gold standard that are not located more than 2Mb from each other) and we counted a predicted interaction as a *true positive* (TP) if it indeed appeared in the gold standard and as a *false positive* if it did not (see Methods for details). We considered predictions at various quality score cut-offs and calculated precision (the proportion of TP in the predicted set) and recall or sensitivity (the proportion of the gold standard that was correctly predicted) at each cut-off value.

Reliable estimation of true and false positive rates requires a large gold standard space (relative to the gold standard itself) and a sufficient overlap between the predictions and the gold standard (space). Figure 4a shows that the gold standard space is indeed ten-fold larger than the gold standard and that after all filtering, 43.34% of the predicted interactions, within the same 2Mb range as the gold standard, lie in the gold standard space, justifying our validation method. For the initial predictions and the predictions after ISMA filtering these percentages are respectively 20.99% and 32.13%.

Figure 4b shows the performance curves, i.e. the precision in function of the recall, of our predicted interactions at different stages of the prediction pipeline. For comparison, we also performed the same validation on the set of long-range interactions reported by Sheffield *et al.* (2013). As expected, the performance of our unfiltered list of interactions and the predictions of Sheffield *et al.* (2013) are comparable since both use the Spearman correlation as a quality score on a comparable number of cell types, the only difference being the filtering of indirect interactions in our method and the limitation to distances less than 500kb by Sheffield *et al.* (2013). However a clear improvement in performance is seen for the filtered interaction lists which result in $\sim$1.5-fold increase in precision at the same level of recall. The same result is observed if we plot precision as a function of the quality score cut-off (Figure 4c).

Because the Hi-C data is itself subject to noise and potentially contains numerous false positive interactions, we used the confidence p-values reported by Jin *et al.* (2013) to construct gold standards from increasingly stringent Hi-C interactions and repeated the same validation experiments. Using the area under the recall precision curve (AUC) as an overall performance measure (Stolovitzky *et al.* (2009)), we found that prediction quality indeed increased signif-

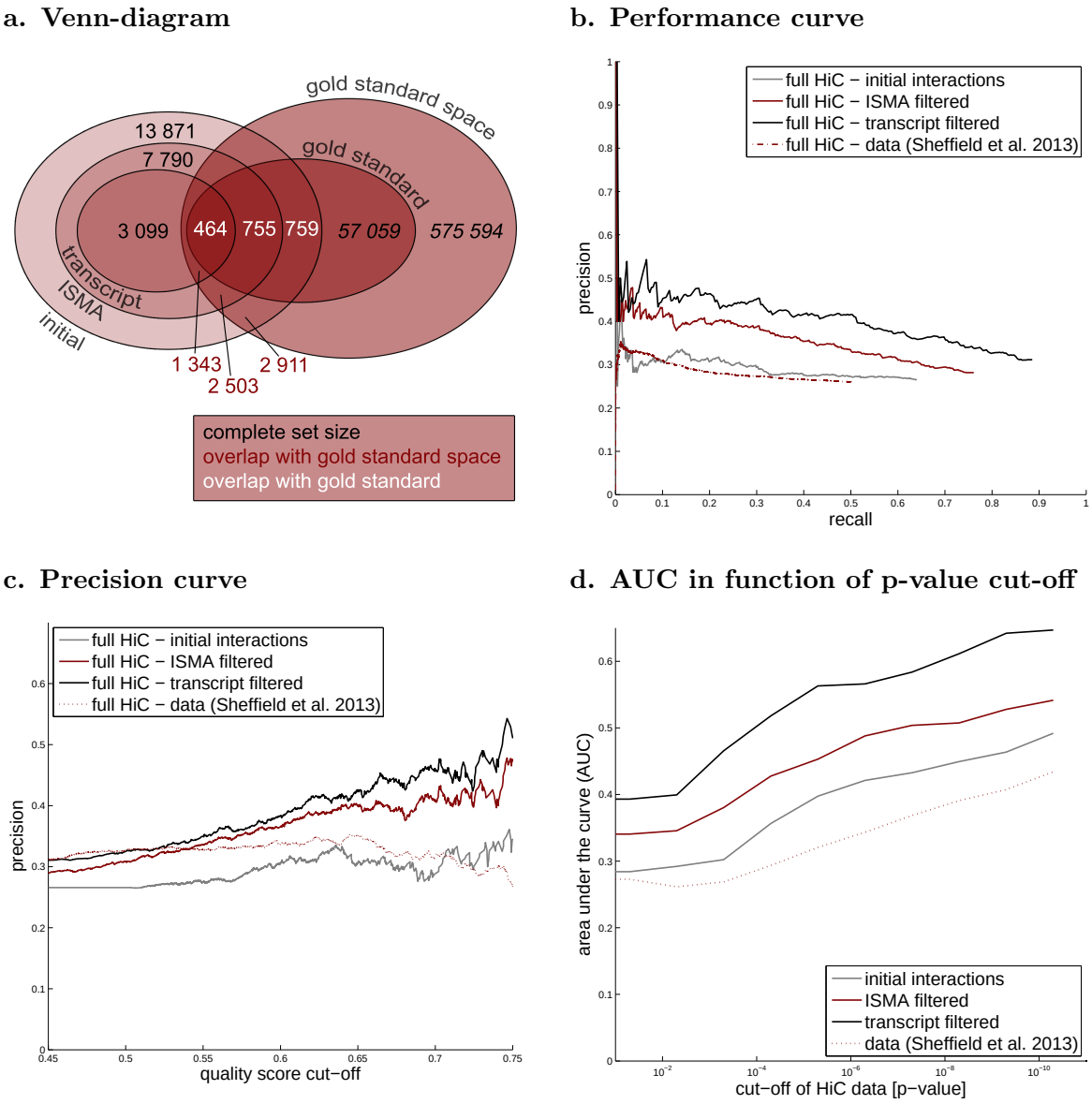


Figure 4: **Validation with Hi-C data.** The gold standard consists of all interactions found in the Hi-C experiments. **a.** To asses the meaningfulness of this validation, it was investigated whether a fair percentage of our predictions are present in the gold standard space. **b.** The performance curve shows the precision in function of the recall for different quality score cut-off values. **c.** The precision curve shows an increasing precision with increasing quality score cut-off values. **d.** The p-value cut-off for the Hi-C data of the gold standard was decreased, resulting in more stringent validation sets. The area under the performance curve (AUC) is shown in function of the p-value cut-off.

icantly for the more stringent sets and that the relative performance of the different filtered sets was conserved across the entire range of stringencies (Figure 4d).

It should be noted that the IMR90 cell type for which the Hi-C data was available was not

part of the cell types from which we predicted the DHS–gene interactions, suggesting that the interactions which could be validated here are not cell-type-specific and that the measured performance is likely an underestimate of the true performance.

2.4.2 Carbon-Copy 3C (5C) validation

The 5C technique makes use of specific primers to identify chromatin interactions. In Sanyal *et al.* (2012) this technique was used to reveal interactions between TSSs of genes and distal elements in 1% of the human genome. These sets of interactions, consisting of 2 sets of primers and 3 cell types, are publicly available in the ENCODE database. From this 5C data we constructed a 'gold standard' (see Methods for details). The limitation to 1% of the genome resulted in only a small overlap between the gold standard space and our predictions (Supplementary Table S6). After all filtering, only 1.65% of our predictions can be validated with this data. Thus a systematic validation using true and false positive rates, similar to the one with the Hi-C data, is not applicable in this case.

We therefore considered each gene that occurred in both the prediction and the validation set separately and counted the number of predicted interactions (# P) and the number of correctly predicted interactions (# C), i.e. that were also found by the 5C technique. Table 2 shows these numbers for the genes that occur in all datasets, i.e. all intermediate predictions and the gold standard (see Supplementary Table S7 for a full list of genes). After all filtering one third of the predicted interactions was positively validated by the 5C data. Furthermore, it is clear that, since the percentages of correctly predicted interactions increase (i.e. 25.30%, 31.25% and 33.33% for respectively the initial, the ISMA filtered and the transcript filtered predictions), the filtering steps indeed improve the quality of the results.

	initial		**ISMA**		**transcript**	
gene	**# P**	**# C**	**# P**	**# C**	**# P**	**# C**
CAV2	30	2	24	2	20	1
CTGF	13	13	12	12	10	10
MAP1A	4	2	4	2	4	2
MET	28	3	18	3	13	3
MOXD1	2	1	2	1	2	1
SELENBP1	2	0	2	0	1	0
SERPINB7	4	0	2	0	1	0
total	**83**	**21**	**64**	**20**	**51**	**17**

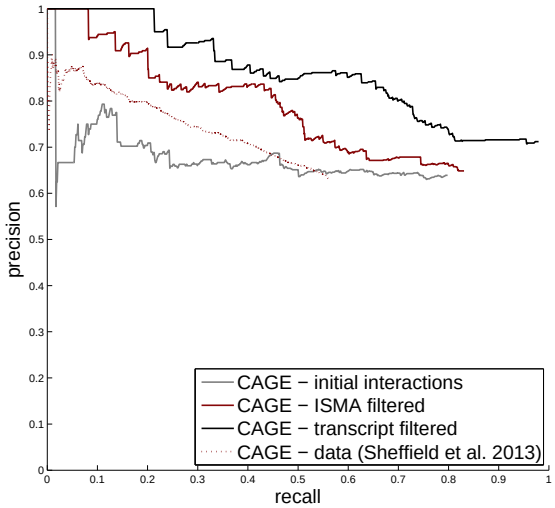
Table 2: **Validation with 5C data.** Only genes that were present in all datasets, i.e. all intermediate results of our calculations and the 5C dataset, are depicted here. # P represents the number of predictions, # C the number of 'correct' predictions with respect to the 5C data. After the first calculation step 25.30% of the interactions are predicted correctly with respect to the 5C validation set. After ISMA filtering this increases to 31.25%, and after transcript filtering to 33.33%.

2.5 Cap Analysis of Gene Expression (CAGE) confirms predicted interactions

CAGE is an experimental technique to identify the promoters and TSS of genes (Shiraki *et al.* (2003)), which has been extensively used within the FANTOM research projects. Recently, in a FANTOM5 research, it has been discovered that, next to small RNA fragments around the TSSs, CAGE also finds small fragments of possible enhancer sites (Andersson *et al.* (2014)). From this data significant enhancer-promoter interactions were predicted based on expression correlation between all pairs of enhancers and promoters within a distance of 500 kbp. This set of predicted interactions was compared with our predictions, as significant overlap between these two sets might assure the validity of both prediction methods.

The validation set now consists of all significant interactions found by the CAGE experiments (Andersson *et al.* (2014)). Similarly to the Hi-C validation, both the performance curves (Figure 5a) and the precision in function of the quality score cut-off values (Figure 5b) were plotted. Figure 5 shows relatively high precision values which indicates a significant overlap between the two datasets. Moreover, precision increases with increasing quality score cut-off values, demonstrating the quality score is a valuable measure to indicate the probability of the interactions. Furthermore, by comparing the curves for the different steps of the calculations, it is clear that the presented method (i.e. combining different datasets and filtering) indeed leads to more accurate predictions.

a. Performance curve



b. Precision curve

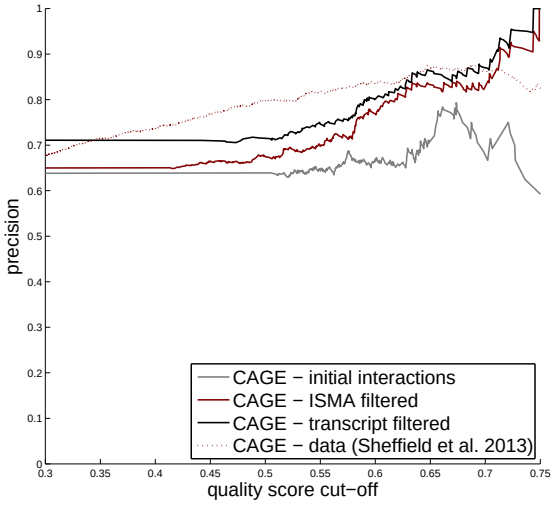


Figure 5: **Validation with CAGE data.** The gold standard consists of all (predicted) interactions from the CAGE experiments. **a.** The performance curve shows the precision in function of the recall for different quality score cut-off values. **b.** The precision curve shows the precision in function of the quality score cut-off.

2.6 Comparison with known long-range interactions

The well-studied *H19/IGF2* locus has been reported to have an imprinted long-range interaction. In the original study (Leighton *et al.* (1995)), carried out on mice, it was demonstrated

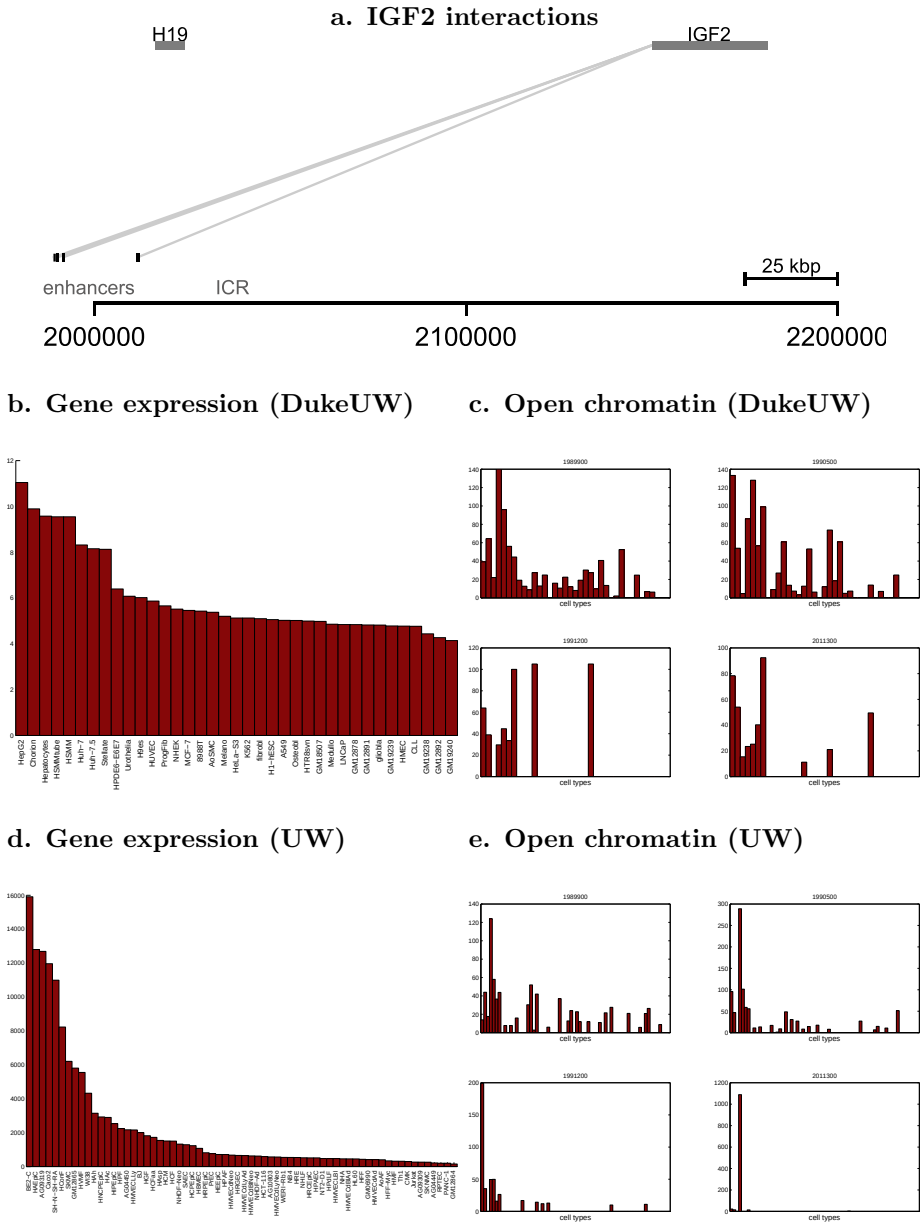


Figure 6: **Predictions show interaction between the *IGF2* gene and 4 DHSs located upstream of the *H19* promoter.** These interactions are confirmed by the resemblance between the gene expression profiles ((b) and (d) for DukeUW and UW respectively) and the open chromatin profiles of the 4 DHSs ((c) and (e) for DukeUW and UW respectively), i.e. higher peaks for the same cell types.

that the *H19* and *IGF2* genes (located on chromosome 7) exhibit parent-of-origin-specific mono-allellic expression. While *H19* is expressed from the maternal chromosome, *IGF2* is expressed from the paternal one. Both genes share enhancer elements and are controlled, besides by the imprinted control region (ICR) situated between the 2 genes, by regions located upstream of the *H19* promoter. This has also been observed in human cell lines (Tabano

et al. (2010)), in which these genes are located on chromosome 11.

Our results show that there is indeed a long-range interaction between the *IGF2* gene and open chromatin regions situated upstream from the *H19* promoter (Figure 6a). This was also reported in Sheffield *et al.* (2013). Although it is reported that these correlations were primarily driven by liver lineages, this long-range interaction is also observed when comparing multiple cell lines.

Figures 6b-e show the gene expression profiles of the *IGF2* gene for the different cell lines of both datasets, together with the DNase sensitivity profiles of the DHSs in question. These profiles show a fair amount of similarity as the higher peaks are observed for the same cell types.

2.7 Exploring the data

To query the predicted interactions, a webservice was developed: http://dhsgen.roslin.ed.ac.uk. By selecting a gene, for which interactions were predicted, from the drop-down list all interacting DHSs can be queried, together with the quality scores of the corresponding interactions. This list of interacting DHSs can be downloaded as a file. In addition, a link to the UCSC Genome Browser is provided in which the interactions are displayed visually.

3 Discussion

One of the principal findings of the ENCODE project has been that regulatory elements occupy a much greater portion of the genome than previously anticipated, but understanding how these elements coordinate the precise spatio-temporal regulation of gene expression remains a formidable challenge. A promising approach to link regulatory elements to their candidate target genes uses guilt-by-association: if the 'activity' (e.g. DNA accessibility or protein-binding frequency) of a regulatory element and expression level of a gene correlate significantly across multiple experimental conditions or cell types, an interaction between them is inferred. Here we improved on existing approaches in two directions. Firstly, we used a subgraph matching algorithm to identify consistent correlations in multiple datasets of matching DNase-seq and exon array samples. This enabled us to incorporate more samples in our analysis while avoiding the need for complex cross-dataset normalization. Secondly, we also considered gene co-expression interactions in our analysis in order to filter interactions between regulatory elements and target genes that are most likely due to indirect effects, again using a subgraph matching approach. This removed the need to limit our search to the area around genes and allowed us to predict interactions across whole chromosomes.

A critical issue when computationally predicting thousands of interactions is to validate them on a correspondingly large scale. We borrowed validation principles from the network reverse-engineering field and showed that there was a significant overlap between our predictions and genome-wide chromosome capture (Hi-C) data as well as predictions derived from CAGE data. This overlap moreover increased when more stringent thresholds were applied to either the predicted interactions or validation data.

Although the various high-throughput experimental technologies used to generate both the training and validation data each have their own biases and limitations, the extent of over-

lapping interactions derived from either of them is highly encouraging and suggests that integrating more data types (e.g. FAIR, ChIP-seq or RNA-seq data) will only lead to more accurate predictions of long-range regulatory interactions. We believe graph-based data integration methods such as the ones introduced here will play a key role in this endeavor.

4 Methods

4.1 Data collection and pre-processing

Both the DNase-seq and the exon array data was collected from the ENCODE database (ENCODE *et al.* (2011)). For all used cell types (see Supplementary Table 1) the DNase-seq peak files were downloaded in bigbed-format and translated to the readable bed-format. Similarly, the exon array data was downloaded for both data sources, i.e. DukeUW (collected data from the Crawford lab of Duke University and the Stamatoyannopoulous lab of the University of Washington) and UW (data from the University of Washington). Where possible, summarized data files were used in which the data of the different experiments is collected. All this data made use of the human genome assembly hg19. This resulted in both DNase-seq and DukeUW exon array data for 37 cell types, and DNase-seq and UW exon array data for 66 different cell types.

The DNase-seq data was preprocessed as follows. The whole genome was divided in 100 bp bins, similar to (Degner *et al.* (2012)), and for each bin B_i a single open chromatin level b_i was calculated as a weighted average of all open chromatin peaks located on this bin:

$$b_i = \sum_j \frac{D_j \cap B_i}{S} d_j$$

in which $(D_j \cap B_i)$ represents the number of overlapping base pairs between the DNase peak D_j and the bin B_i, S is the bin size (in this case 100 bp) and d_j represent the open chromatin level of DNase peak D_j.

Subsequently, for each chromosome (except the sex chromosomes) and for each data source, a gene expression matrix and an open chromatin matrix were generated with respectively the gene expression and the open chromatin levels. The gene expression matrix is a $M \times N$ matrix, with M the number of cell types in the data source and N the number of genes/exons in the chromosome. The open chromatin matrix is a $M \times K$ matrix, with M as defined previously and K the number of bins in the chromosome. Supplementary table S5 shows for each chromosome the dimensions of these matrices.

4.2 Calculating correlations

For each chromosome and for each data source, absolute Spearman correlations were calculated between the columns of the open chromatin matrix and the columns of the gene expression matrix. To limit the calculations only the genes/exons for which the expression levels vary sufficiently across the different cell types, i.e. genes g_i for which $\text{std}(g_i) > \text{mean}_{n=1..N}(\text{std}(g_n))$, are taken into account, resulting in a $M \times N'$ gene expression matrix. Similarly, only those

bins that have open chromatin in at least one of the cell types, are taken into account, resulting in a $M \times K'$ open chromatin matrix. Calculating the correlations results in a $K' \times N'$ correlation matrix (See Supplementary Table S5 for the values of N' and K').

Subsequently, the interactions with the most significant correlations were selected by applying a false discovery rate (FDR) threshold of 10%. For this, the absolute Spearman correlations between a random open chromatin matrix (i.e. the open chromatin matrix with the rows permuted randomly) and the gene expression matrix were calculated n times (in our case $n = 3$). Then a correlation cut-off value was determined, so that only 10% of the remaining correlations is due to randomness, i.e. applying this cut-off value to the random correlation matrix results in only 10% of the interactions in comparison to applying this cut-off value to the original correlation matrix. All interactions with a significant correlation, i.e. higher than the determined cut-off value, were then kept in a list in which each entry consists of a gene/exon, an open chromatin area and a quality score, represented by the absolute correlation value.

4.3 ISMA Filtering

A weighted edge-colored graph was built form the resulting interactions of phase 1 for each dataset, together with location information of the genes/exons of each dataset. All interactions are translated to edges between open chromatin bins and genes/exons, in which each data source is represented by a different edge type (i.e. color). Moreover, edges are created between genes and exons that are co-located, i.e. alignment links. Weights were assigned to all edges. While for the interaction links this is the correlation value, for the alignment links this is a unit weight, i.e. 1.

Subsequently, the Index-based Subgraph Matching Algorithm (ISMA) (Demeyer *et al.* (2013)) was applied in this graph to find all subgraphs representing a significant interaction between a bin and a gene/exon that occurs in both datasets. These subgraphs (see Figure 1b) are 3-node graphs with 3 different edge types: an interaction between a bin and a gene (dataset 1), an interaction between a bin and an exon (dataset 2) and an alignment between a gene and an exon. When applying the ISMA algorithm a score was calculated for each subgraph by multiplying all edge weights. The square root of this score, which is the geometric mean of the two correlations, is then used as the quality score of the interaction.

4.4 Clustering and Transcript Filtering

The motifs (i.e. subgraphs) found by the ISMA algorithm are clustered in order to find graph structures similar to the one depicted in figure 1c. All subgraphs with the same open chromatin region and the same gene are collected together. Each of this clusters represents a single interaction between a DHS and a gene.

Subsequently, this list of clusters is filtered making use of the publicly available A-MEXP-2246 annotation file[1] to only keep those clusters that represent known gene transcripts. Only those interactions are retained for which the set of exons in the cluster contains the majority (i.e. $> 80\%$) of exons of a known gene transcript.

[1] http://www.ebi.ac.uk/arrayexpress/files/A-MEXP-2246/A-MEXP-2246.additional.1.zip

Finally, the quality score of the interaction was calculated as the maximum quality score of all motifs (i.e. 3-node subgraphs) participating in the corresponding cluster.

4.5 Indirect Filtering

To eliminate indirect interactions (see Figure 2a), the absolute Spearman correlation was calculated between each pair of genes/exons. Subsequently, a network was constructed with the set of nodes consisting of all DHSs and genes/exons and two types of links: interaction links between DHSs and genes/exons and correlations between genes/exons. The ISMA algorithm was again applied to enumerate all subgraphs with a configuration as depicted in figure 2b. Subsequently, in each of these subgraphs the link with the lowest weight was identified and if this was an interaction link, the corresponding interaction was removed from the predicted interactions.

Next to this, after all filtering, interactions between DHSs and co-located genes were removed from the predictions.

4.6 Validation with Hi-C and CAGE data

Firstly, both the Hi-C and the CAGE validation set were retrieved from their corresponding publications (Jin *et al.* (2013); Andersson *et al.* (2014)).

For fair comparison, our set of predicted interactions was limited to only those in the same distance range and both sets were restricted to only those interactions between open chromatin areas and genes which occur in both sets (similar to Stolovitzky *et al.* (2009)). Moreover, the open chromatin areas were translated to the same level of detail. While we opted for 100 bp bins to represent the open chromatin areas, the validation sets utilize different open chromatin regions. The most detailed information is translated to the least detailed. For example, the regions of the Hi-C data contain on average 17,540 bp. This means that each of our DHS bins needed to be translated to the region (of the Hi-C data) in which it is located.

Subsequently, for different quality score cut-off values the set of predicted interactions was compared with the validation set and the precision and recall were calculated as follows:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

with TP the number of true positives, FP the number of false positives and FN the number of false negatives. Plotting the precision in function of the recall for different quality score cut-off values results in the performance curve. Similarly, to investigate the avail of the quality score the precision was plotted in function of the quality score cut-off.

4.7 Validation with 5C data

The 5C data, as published in (Sanyal *et al.* (2012)), was collected from the ENCODE database. This dataset consists of 5C data for 3 cell types (GM12878, K562 and HeLa-S3) and 2 sets of

primers. For each set of primers, all interactions of the different cell types were collected and those interactions that are present in at least 2 cell types were added to the gold standard.

Subsequently, for each gene we counted the number of predicted interactions and the number of 'correctly' predicted interactions. Hereby, we limited the set of predictions to only those interactions of which both the gene and the DHS are present in the gold standard.

5 Data Access

All the data used in this research is publicly available. Both the exon array and the DNase-seq data were collected from the ENCODE database. Both the Hi-C and the CAGE validation data was gathered from the corresponding publications. The 5C validation set was collected from the ENCODE database. The predicted interactions can be queried at http://dhsgen.roslin.ed.ac.uk.

6 Acknowledgements

We thank David Hume for providing us with early access to the CAGE data and results. We thank Andy Law for the help with the webservice. This research was supported by Roslin Institute Strategic Grant funding from the BBSRC.

7 Figures

List of Figures

8 Tables

List of Tables

References

Akalin A, Fredman D, Arner E, Dong X, Bryne J, Suzuki H, Daub C, Hayashizaki Y, Lenhard B (2009) Transcriptional features of genomic regulatory blocks. *Genome Biology* **10**: R38

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, Ntini E, Arner E, Valen E, Li K, Schwarzfischer L, Glatz D, Raithel J, Lilje B, Rapin N, Bagger FO, *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature* **507**: 455–461

Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**: 1074–1077

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**: 311–322

de Wit E, de Laat W (2012) A decade of 3C technologies: insights into nuclear organization. *Genes & Development* **26**: 11–24

Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, *et al.* (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**: 390–394

Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing chromosome conformation. *Science* **295**: 1306–1311

Demeyer S, Michoel T, Fostier J, Audenaert P, Pickavet M, Demeester P (2013) The index-based subgraph matching algorithm (ISMA): fast subgraph enumeration in large networks using optimized search trees. *PloS One* **8**: e61183

Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, *et al.* (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Research* **16**: 1299–1309

ENCODE PC, *et al.* (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biology* **9**: e1001046

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**: 43–49

Galas DJ, Schmitz A (1978) DNase footprinting a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Research* **5**: 3157–3170

Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen CA, Schmitt AD, Espinoza CA, Ren B (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**: 209–294

Kapur K, Xing Y, Ouyang Z, Wong WH (2007) Exon arrays provide accurate assessments of gene expression. *Genome Biology* **8**: R82

Leighton PA, Saam JR, Ingram RS, Stewart CL, Tilghman SM (1995) An enhancer deletion affects both H19 and Igf2 expression. *Genes & development* **9**: 2079–2089

Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, *et al.* (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**: 84–98

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289–293

Marstrand TT, Storey JD (2014) Identifying and mapping cell-type-specific chromatin programming of gene expression. *Proceedings of the National Academy of Sciences* **111**: E645–E654

Michoel T, Nachtergaele B (2012) Alignment and integration of complex networks by hypergraph-based spectral clustering. *Physical Review E* **86**: 056111

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**: 1344–1349

Natarajan A, Yardımcı GG, Sheffield NC, Crawford GE, Ohler U (2012) Predicting cell-type–specific gene expression from regions of open chromatin. *Genome Research* **22**: 1711–1722

Noonan JP, McCallion AS (2010) Genomics of long-range regulatory elements. *Annual Review of Genomics and Human Genetics* **11**: 1–23

Sanyal A, Lajoie BR, Jain G, Dekker J (2012) The long-range interaction landscape of gene promoters. *Nature* **489**: 109–113

Sheffield NC, Thurman RE, Song L, Safi A, Stamatoyannopoulos JA, Lenhard B, Crawford GE, Furey TS (2013) Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome research* **23**: 777–788

Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, *et al.* (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences* **100**: 15776–15781

Stolovitzky G, Prill RJ, Califano A (2009) Lessons from the DREAM2 Challenges. *Annals of the New York Academy of Sciences* **1158**: 159–195

Tabano S, Colapietro P, Cetin I, Grati FR, Zanutto S, Mandò C, Antonazzo P, Pileri P, Rossella F, Larizza L, *et al.* (2010) Epigenetic modulation of the IGF2/H19 imprinted domain in human embryonic and extra-embryonic compartments and its possible role in fetal growth restriction. *Epigenetics* **5**: 313–324

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature* **489**: 75–82

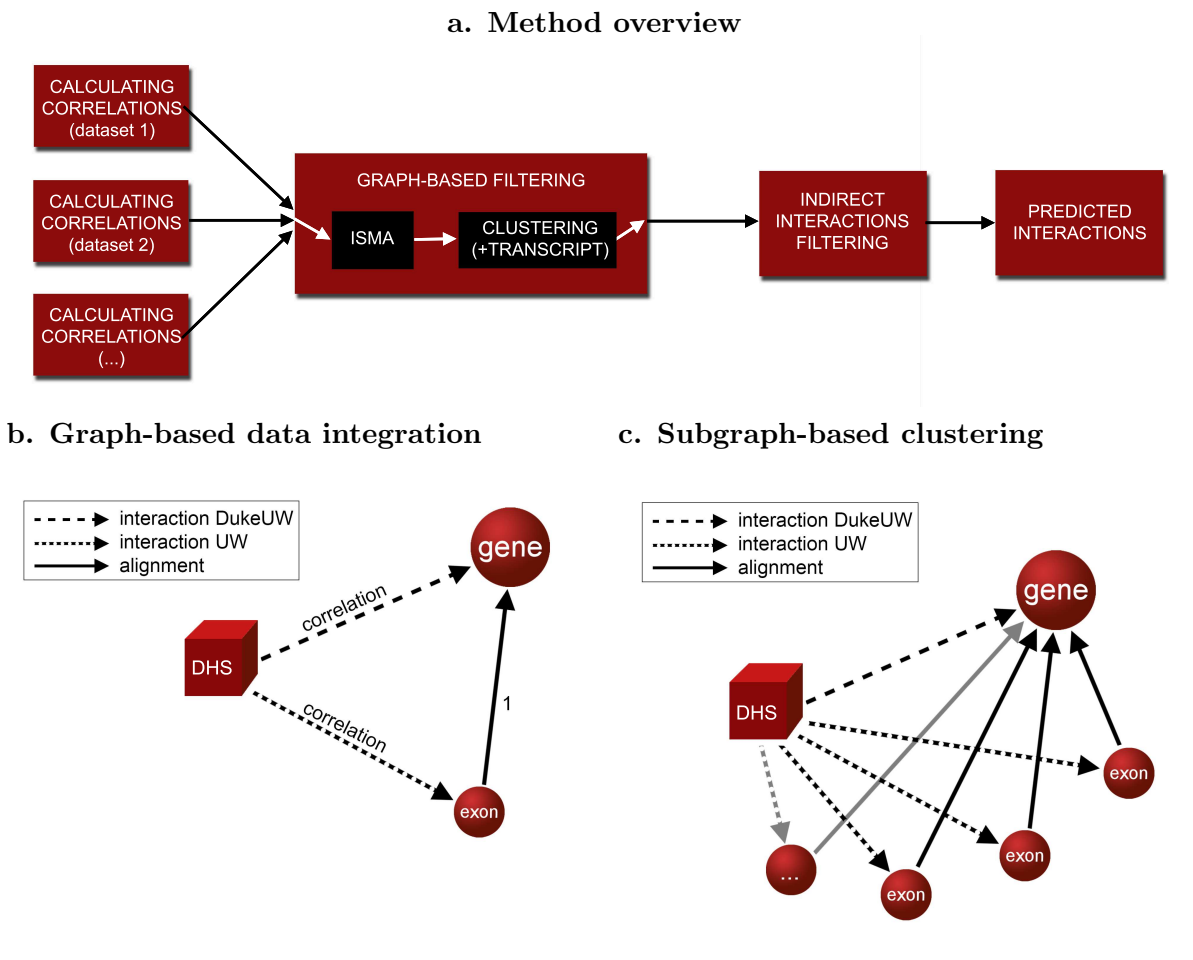
Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W (2002) Looping and interaction between hypersensitive sites in the active β-globin locus. *Molecular Cell* **10**: 1453–1465

Wei G, Zhao K (2011) 3C-based methods to detect long-range chromatin interactions. *Frontiers in Biology* **6**: 76–81

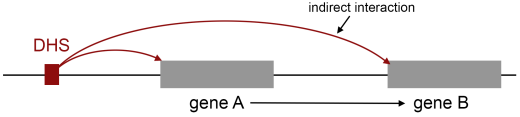
Wu C (1980) The 5'ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. *Nature* **286**: 854–860

Xi H, Shulha HP, Lin JM, Vales TR, Fu Y, Bodine DM, McKay RD, Chenoweth JG, Tesar PJ, Furey TS, *et al.* (2007) Identification and characterization of cell type–specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genetics* **3**: e136

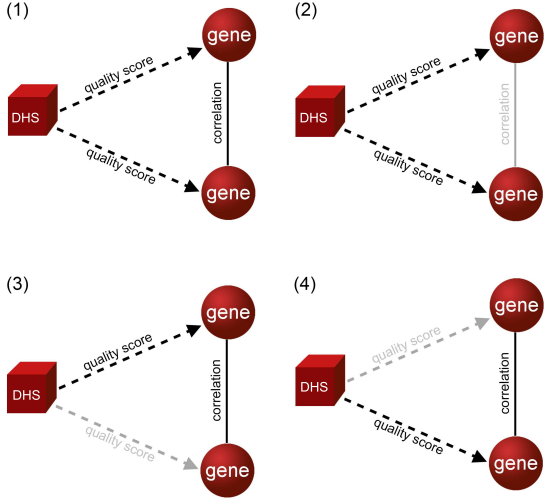
Zhao Z, Tavoosidana G, Sjölinder M, Göndör A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U, *et al.* (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra-and interchromosomal interactions. *Nature Genetics* **38**: 1341–1347


a. Method overview
CALCULATING CORRELATIONS (dataset 1)
CALCULATING CORRELATIONS (dataset 2)
CALCULATING CORRELATIONS (...)
GRAPH-BASED FILTERING
ISMA
CLUSTERING (+TRANSCRIPT)
INDIRECT INTERACTIONS FILTERING
PREDICTED INTERACTIONS
b. Graph-based data integration
interaction DukeUW
interaction UW
alignment
gene
DHS
correlation
correlation
1
exon
c. Subgraph-based clustering
interaction DukeUW
interaction UW
alignment
gene
DHS
exon
exon
exon
...

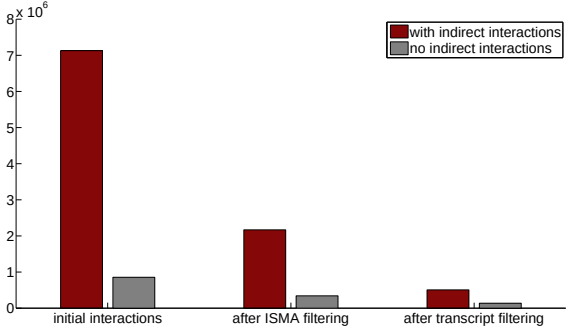
a. Indirect DHS–gene interactions



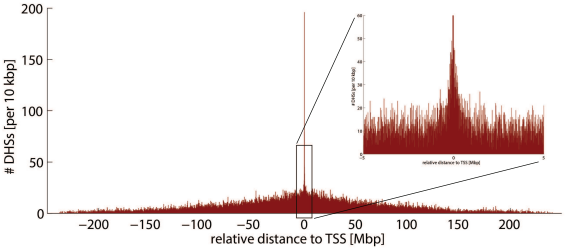
b. Graph-based filtering



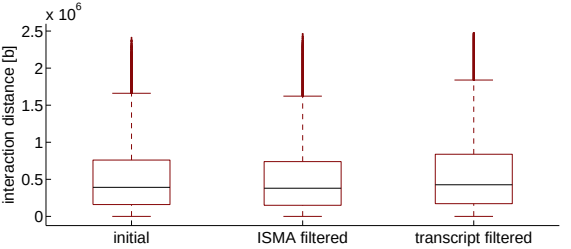
c. Interaction numbers



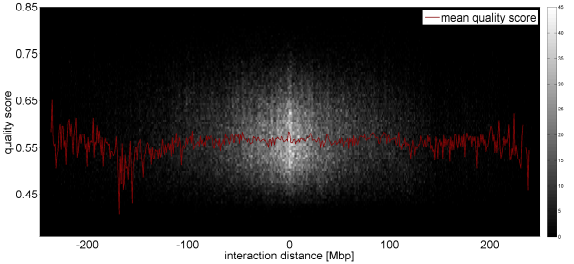
a. Number of DHSs in function of interaction distance



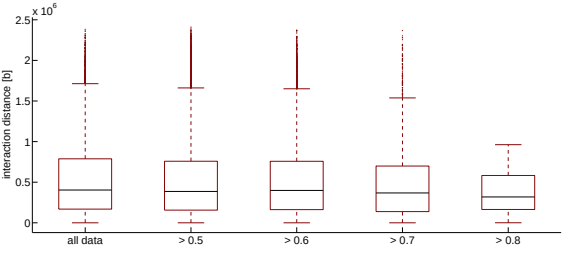
b. Distribution of interactions distances



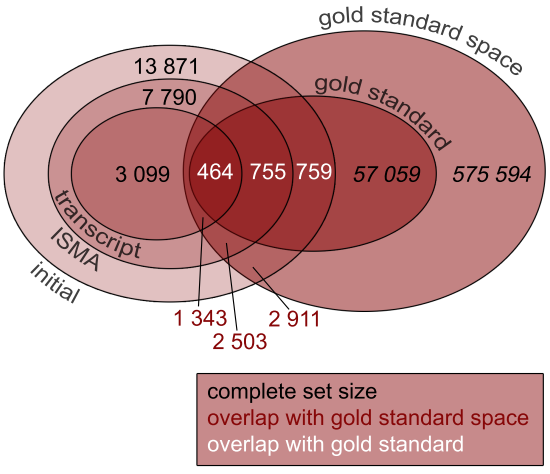
c. Quality score in function of interaction distance



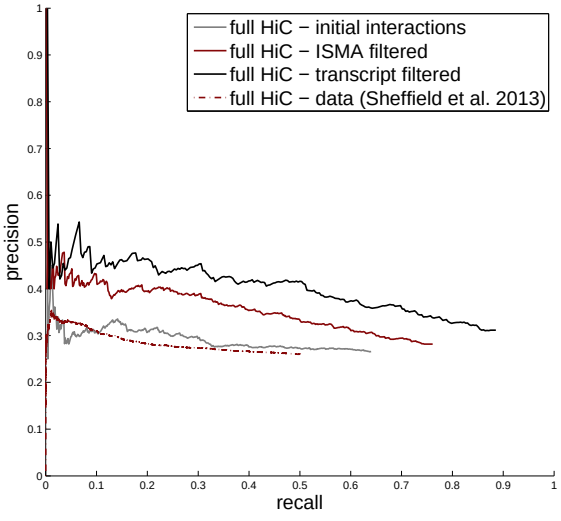
d. Distribution of interaction distance in function of quality score range



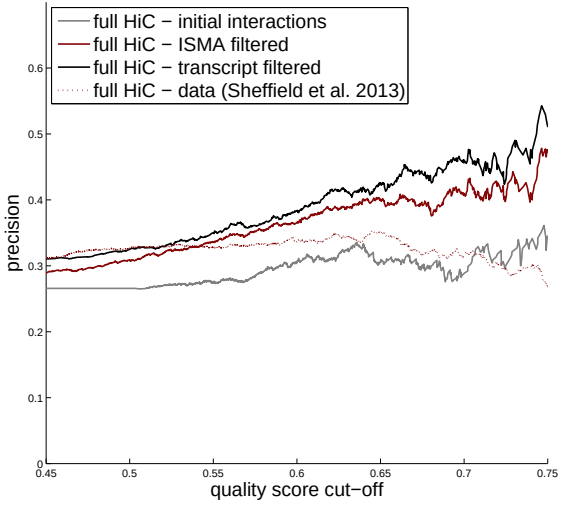
a. Venn-diagram



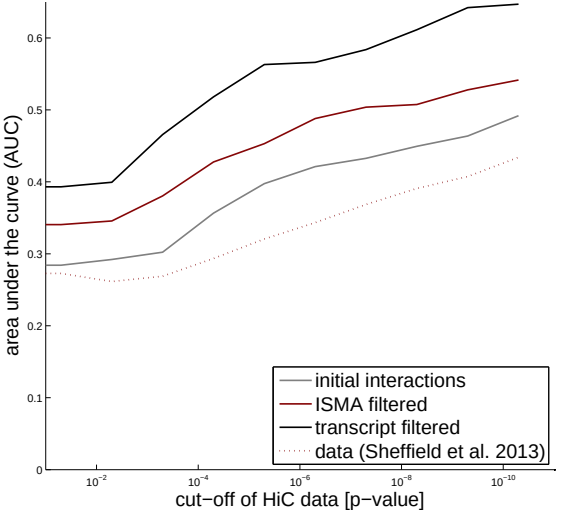
b. Performance curve



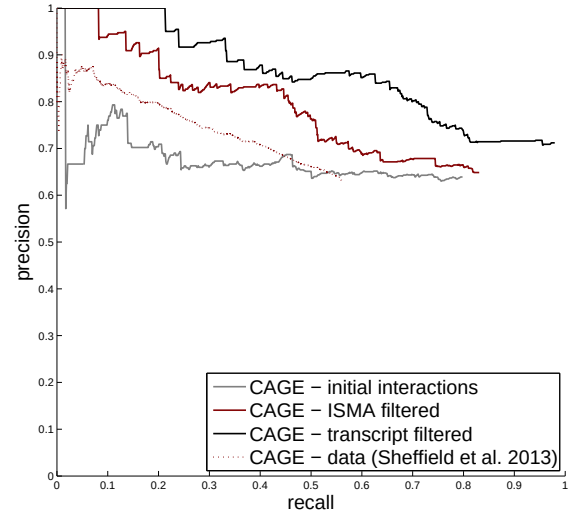
c. Precision curve



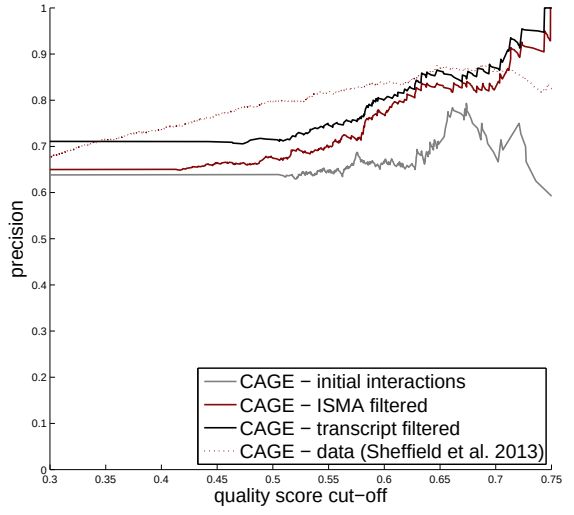
d. AUC in function of p-value cut-off



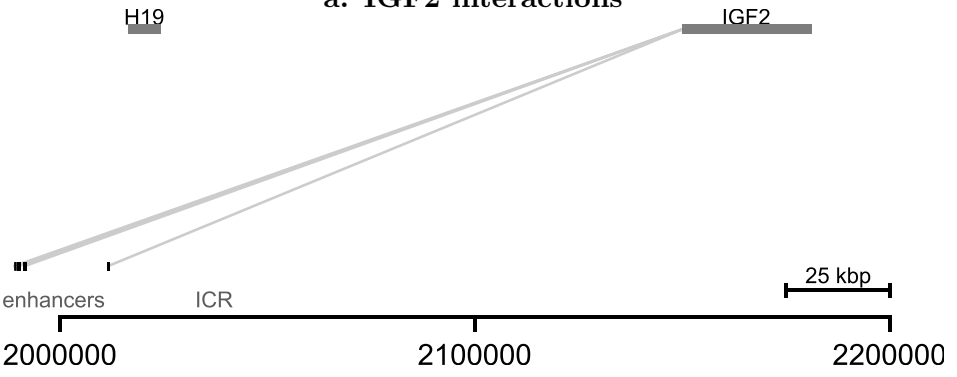
a. Performance curve



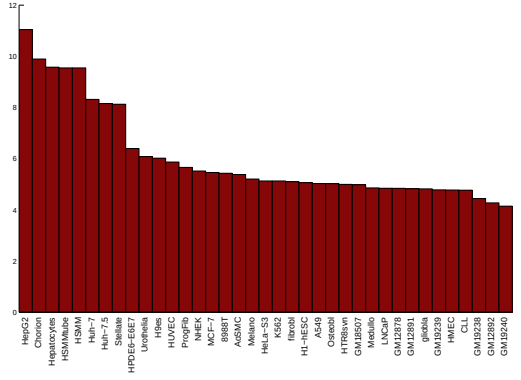
b. Precision curve



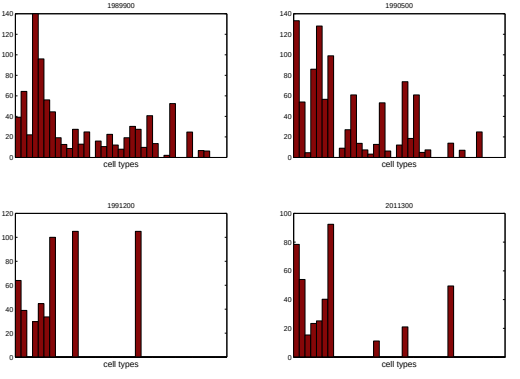
a. IGF2 interactions



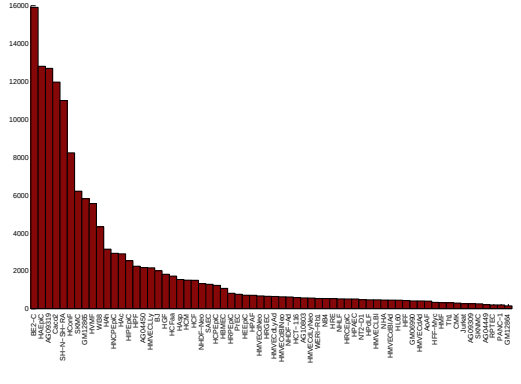
b. Gene expression (DukeUW)



c. Open chromatin (DukeUW)



d. Gene expression (UW)



e. Open chromatin (UW)